DOCUMENT RESUME

ED 413 755 FL 024 776

AUTHOR Kruyt, J. G.; Raaijmakers, S. A.; van der Kamp, P. H. J.; van Strien, R. J.

TITLE On-Line Access to Linguistically Annotated Text Corpora of Dutch via Internet.

PUB DATE 1995-00-00

NOTE 7p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Access to Information; *Computational Linguistics; Computer Software; *Discourse Analysis; *Dutch; Foreign Countries; Information Retrieval; *Internet; *Language Research; Lexicology; Linguistic Theory; Newspapers; Uncommonly Taught Languages

IDENTIFIERS *Language Corpora; Netherlands

ABSTRACT

Corpora of present-day Dutch developed by the Institute for Dutch Lexicology include two linguistically annotated corpora that can be accessed via Internet: a 5-million word corpus covering a variety of topics and text types, and a 27-million word newspaper corpus. The texts of both were acquired in machine-readable form and have been lemmatized and tagged and loaded onto an online retrieval system. Queries may address the entire corpus or a subcorpus defined by the user. The present user interface appears complex, particularly for inexperienced users, due to a high degree of formalism, but efforts are being made to reduce formalism. A prototypical natural language interpreter is under development. Copyright restrictions limit the transfer of information to the user's electronic mail. Access to the corpora is free for non-commercial research purposes with a signed personal user agreement. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 755

On-line Access to Linguistically Annotated Text Corpora of Dutch via Internet

J.G. Kruyt, S.A. Raaijmakers, P.H.J. van der Kamp, R.J. van Strien*

Institute for Dutch Lexicology INL
P.O. Box 9515
NL-2300 RA Leiden
Tel.: +31 71 527 2270
Fax: +31 71 527 2115
E-mail: kruyt@ruLxho.Leidenuniv.nl

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Norbert Vd3
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

FL024796

* The first two authors are working at the linguistic Language Database Department, the other two at the Electronic Data Processing Department.

1. Corpus development at the Institute for Dutch Lexicology INL

The Institute for Dutch Lexicology INL is a research institute subsidized by the Dutch and Belgian governments. Corpus development at INL dates from the mid-seventies. Up to 1990, the INL text corpora were developed for lexicographical purposes mainly. Presently, they are used for a broad range of research and applications (cf. Van Sterkenburg and Kruyt in press). A recent example is the official Dutch spelling guide published in 1995, which is based on INL text corpora (Kruyt and Van Sterkenburg this volume).

INL text corpora of present-day Dutch include two linguistically annotated corpora which can be consulted via the international computer network Internet: the *5 Million Words Corpus 1994*, which covers a variety of topics and text types, and the *27 Million Words Newspaper Corpus 1995*. A corpus of ca. 30 million words, with varied composition and with extended linguistic encoding, will be ready for similar use in spring 1996. The present paper reports on the former two corpora already accessible via Internet.

2. Characteristics of the corpora

The *5 Million Words Corpus 1994* contains seventeen text sources, most of them dating from 1989-1994. The texts are classified along the parameters publication medium (book, newspaper, magazine, written-to-be-spoken) and topic (politics, journalism, leisure, linguistics, environment, business and employment). The *27 Million Words Newspaper Corpus 1995* covers one newspaper only, with editions dating from 1994 and 1995.

The texts of both corpora were acquired in machine-readable form, on a contract basis with the provider. The contract specifies the conditions of use, taking into account issues of copyright. Permission has been obtained for use of the texts in this particular application. After some preprocessing (Kruyt and Van Sterkenburg this volume), the texts were input for automatic linguistic encoding. Part of speech (POS) and headword were automatically assigned to the word forms in the electronic texts by lemmatizer/POS-taggers developed by INL. The lemmatizer/POS-tagger DutchTale (Van der Voort van der Kleij et al. 1994) was applied to the *5 Million Words Corpus 1994*. An improved version of this program has been used for encoding the *27 Million Words Corpus 1995*. This new version, *DutchTale II*, uses separate rule files, which allows for easy inspection and modification of the implemented linguistic knowledge. The addition of a more elaborate morphological module, incorporating, amongst others, compound analysis, has resulted in an increased number of analysable

tokens (individual word forms). Supplementary disambiguation rules have contributed to a higher precision of disambiguation. *DutchTale II* is implemented in C and runs on the institutional VAX.

Most of the data has not been corrected, neither at the level of the proper text, nor at the level of POS and headword.

3. Retrieval facilities

The linguistically encoded texts were loaded into an on-line retrieval system developed by INL. Queries may address the whole corpus, or a subcorpus defined by the user. Parameters for the definition of subcorpora are text source, topic and publication medium for the *5 Million Words Corpus 1994*, and year and month of publication for the *27 Million Words Newspaper Corpus 1995*. The system allows the user to search for single words or word patterns, including some, rather primitive, predefined syntactic patterns which can be customized by the user. Search definitions may include references to word forms, POS and headwords, both separately and in combination by use of Boolean operators and proximity searches. Some examples of queries are:

(Boolean) lemma='hongar*' and not pos='a'

This query searches for lemmas compliant with the pattern 'hongar*' (the asterisk serves as a wildcard) with part of speech not equal to 'a' (adjective).

(proximity search) lemma='president+koning*+staatshoofd'][?|0..3]<'PP'>

In this query the '+' acts as the Boolean operator OR. So, the query searches for lemmas compliant with either 'president' OR 'koning*' OR 'staatshoofd' followed by a 'PP' (prepositional phrase) within at most 3 arbitrary words.

The present user interface appears to be rather complex, in particular for unexperienced users, due to the high degree of formalism. During the seminar, a more elegant user interface was demonstrated, containing a reduced-formalism interpreter. The interpreter allows the user to enter his query with a less elaborate notation. The retrieval engine, however, works with the complex formalism, so translation is necessary. With this interface, the latter example can be entered as:
le=president or koning* or staatshoofd dist 3 ? cat=PP

Also, a prototypical natural language interpreter is under development. This interpreter accepts queries in plain Dutch. An example is:

geef mij alle lemmas die niet op "heid" uitgaan
'give me all lemmas that do not end with "heid"'
which is translated into: not lemma='*heid'

The natural language interpreter will operate in tandem with the reduced-formalism interpreter; if the natural language interpreter fails to comprehend the query, the user will have to address the reduced-formalism interpreter. This interface will be implemented in the *30 Million Words Corpus* planned for 1996.

Output data of the retrieval system include intermediate tables with the possibility of selecting specific items (word forms, lemmas and POS with their frequencies), and ultimately a series of concordances of the searched item(s) (i.e. the searched term(s) in the local context), with a user-defined context size. Concordances can be sorted by the user along several parameters. A few concordances for the Boolean and proximity searches formulated above are:

27 Million Words Newspaper Corpus 1995
For the INL, Leiden 11 16 1995, version 1.01

NRC_NOV_94* terracotta vazen uit	**Hongarije.**	„Ik heb een grote boerderij, di
NRC_NOV_94* s collega's uit Polen,	**Hongarije,**	Tsjechië, Slowakije, Roemenië en
NRC_NOV_94* se diplomaat. Polen en	**Hongarije,**	die al formeel het lidmaatschap
NRC_NOV_94* ehouden met Estland en	**Hongarije,**	maar daar heb ik nooit het predi
NRC_NOV_94* munistische landen als	**Hongarije**	en Bulgarije. De grondwet definie
NRC_NOV_94* ë, Slowakije, Polen en	**Hongarije.**	Terugkijkend waren er al in sept

...

<PREV>/<NEXT>=previous/next page, <8/HELP>=help

27 Million Words Newspaper Corpus 1995
For the INL, Leiden 11 16 1995, version 1.01

NRC_NOV_94* es van commissaris der	**koningin in Zuid-Holland** en secretaris-gene
NRC_NOV_94* e. Daarvoor zou men de	**president van de Europese Beweging** in Frank
NRC_NOV_94* ordeel Jakarta, 1 Nov.	**President Soeharto van Indonesië** acht de ad
NRC_NOV_94* aangespannen tegen het	**staatshoofd wegens onbehoorlijk bestuur.** He
NRC_NOV_94* Hafr Al-Baten, 1 Nov.	**Koning Fahd van Saoedi-Arabië** heeft toegege
NRC_NOV_94* i Boldyrev, dat hij de	**president herhaaldelijk heeft ingelicht** ove
NRC_NOV_94* lag. Daarop verdedigde	**president Jeltsin Gratsjov in zeer lovende**
NRC_NOV_94* en woordvoerder van de	**president in Kaapstad** ondubbelzinnig had on

...

<PREV>/<NEXT>=previous/next page, <8/HELP>=help

Due to copyright restrictions, a limited number of concordances can be transferred to the user's computer by e-mail. It is not allowed to transfer complete texts or substantial text fragments.

The retrieval system is running on a VAXstation 4000/90A, under OpenVMS 6.0. It was developed in VAX Pascal, using the VAX SMG-routines for screen handling (cf. Van der Voort van der Kleij et al. 1994). The more elegant user interface was developed with TPU, a user-extendible text processor for VAX systems. The INL VAXstation is a multi-user computer concurrently used by many colleagues working on various INL projects. In order to restrain the guest users from accessing data to which they are not authorized, they are locked up in so-called captive accounts, a feature of the VAX/OpenVMS operating system. These accounts allow them only to run the retrieval program(s) for which they have signed the corresponding user agreement(s) (see below). Furthermore, all actions of the guest users are stored in logfiles, which are used for internal statistics and security reports. Additionally, an analysis of the list of queries will be used for enhancing the retrieval system.

4. Access to the corpora.

Consulting the corpus is free of charge for non-commercial, research purposes, provided that a personal user agreement is signed. The user agreement includes the conditions of use. For academic teaching purposes, special arrangements are possible after consultation with the first author or the director of INL, Prof. dr. P.G.J. van Sterkenburg. The conditions for commercial applications are to be discussed with the director of INL.

To gain access to the corpus, an electronic user agreement form is to be obtained from our mailserver Mailserv@Rulxho.Leidenuniv.NL. Type in the body of your e-mail message: SEND [5MLN94]AGREEMNT.USE or SEND [27MLN95]AGREEMNT.USE, for the *5 Million Words Corpus 1994* and the *27 Million Words Newspaper Corpus 1995*, respectively. Please make a hard copy of the agreement form, sign it, keep a copy yourself, and return a signed copy to: Institute for Dutch Lexicology INL, P.O. Box 9515, 2300 RA Leiden. After receipt of the signed user agreement, you will be informed about your username and password. Note that the use of a VT 220 (or higher) terminal, or an appropriate terminal-emulator (e.g. Kermit) is recommended. If you need additional information, please send an e-mail message to Helpdesk@Rulxho.Leidenuniv.NL.

References

Kruyt, J.G. and P.G.J. van Sterkenburg. "A New Dutch Spelling Guide". This volume.

Sterkenburg, P.G.J. van and J.G. Kruyt. In press. "Dutch Electronic Corpora: their history, applications and future". *Computers and the Humanities.*

Voort van der Kleij, J.J. van der, S. Raaijmakers, M. Panhuijsen, M. Meijering and R. van Sterkenburg. 1994. "Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem". Noordman, L.G.M. and W.A.M. de Vroomen (eds.). *Informatiewetenschap 1994. Wetenschappelijke bijdragen aan de derde STINFON-conferentie,* Tilburg, 181–194.